DOCUMENT RESUME

ED 306 249 TM 013 027

AUTHOR Ferrara, Steven; Walker-Bartnick, Leslie
TITLE Constructing an Essay Prompt Bank Using the Partial Credit Model.
PUB DATE Mar 89
NOTE 39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 29, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cues; Equated Scores; *Essay Tests; Grade 9; High School Freshmen; *Item Banks; Latent Trait Theory; *Prompting; Scaling; Secondary Education; State Programs; *Test Construction; Testing Programs; Writing (Composition); *Writing Evaluation
IDENTIFIERS Maryland Writing Test; *Partial Credit Model; Rasch Model; Writing Tests

ABSTRACT

The viability of the Partial Credit Model (PCM) of G. N. Masters (1982) for scaling and banking individual essay prompts from a statewide writing assessment program was investigated. Preliminary analyses indicated the feasibility of using the PCM, the general polychotomous form of the Rasch model, with essay data from the Maryland Writing Test (MWT). Subjects were a cohort of ninth graders taking the MWT in the years 1984 through 1988. Data were the averaged narrative and explanatory scores from the administrations of the MWT in these years with their operative prompts and from the administration of the 1984 prompts used as anchor prompts each year in an equating study. Data were analyzed with the MSTEPS program of B. D. Wright et. al. (1988). Equipercentile and PCM banking procedures agreed on relative test difficulty in these data sets. Because of positive results from the analysis of link consistency among prompts and comparisons between equipercentile and PCM "equatings," the PCM appears viable for banking MWT prompts. Benefits of a prompt bank would include simplification of the prompt selection and evaluation process. Seven tables and two graphs present study data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306249

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEVEN FERRARA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Constructing an Essay Prompt Bank Using the Partial Credit Model

Steven Ferrara
Leslie Walker-Bartnick

Maryland State Department of Education

Paper presented at the annual meeting of the National Council on Measurement in Education, March 29, 1989, San Francisco

TM013027

BEST COPY AVAILABLE

Constructing an Essay Prompt Bank Using the Partial Credit Model

Steven Ferrara
Leslie Walker-Bartnick

Maryland State Department of Education
March 15, 1989

The purpose of this study is to test and demonstrate the viability of using the Partial Credit model of Masters (1982) for scaling and banking individual essay prompts from a statewide writing assessment program. Producing annual forms of direct writing assessments that are equivalent in difficulty and, therefore, produce scores that are expressible on the same underlying score scale, has been a perennial problem for statewide and other writing competency testing programs. Test form and score equivalence is especially important to direct writing assessments like the Maryland Writing Test (MWT), in which students must pass a test comprised of two essays in order to receive a high school diploma, and to other writing assessment programs intended to demonstrate growth in writing performance across years. In this study we examine the viability of the Partial Credit model for creating an essay prompt bank using existing data; if the results suggest that the model may be viable with these data, it would allow us to use the model to create a bank of new, calibrated prompts for use in future test administrations.

Typical procedures for producing test forms of equivalent difficulty and that produce approximately equivalent scores usually involve several steps: (1) applying standardized guidelines for generating new prompts; (2) selecting for operational administration only those individual prompts and prompt pairs that produce score distribution locations and shapes similar to anchor prompt distributions; (3) developing training and benchmark materials consistent with previous years' materials; (4) training experienced raters using explicit scoring rules, criteria, and benchmark essays; (5) monitoring and correcting raters' performance during scoring; and (6) conducting follow-up studies of the degree to which scoring criteria are applied consistently from year to year (see Goldberg & Walker-Bartnick, 1988). In some cases, ex post facto statistical equating is also used to equate scores (see Ferrara, 1987; Phillips, 1987). However, for the Maryland Functional Testing Program it would also be useful to be able to (1) select individual prompts -- not just intact pairs of prompts -- for operational administration, (2) pre-equate total test scores (Williams & Phillips, 1984) so that post-equating score adjustments are unnecessary, and (3) provide a statistical link between annual test editions and an anchored score scale to support procedural links. Using the Partial Credit model to create prompt banks for

the MWT and other essay testing programs should allow realization of these benefits and the other usual benefits of item response theory models.

The Partial Credit Model

The Partial Credit (PC) model is the general polychotomous form of the Rasch model. It is designed for use with personality, attitude, educational achievement, and other items that elicit responses that are scored in ordered categories; that is, for item responses that can be awarded partial credit and not scored only correct/incorrect. Unlike the Rasch model for dichotomous items, items calibrated with the PC model have associated with them several difficulty values, each indicating the difficulty of reaching a "step" along the way to successful completion of an item. Like the Rasch model for dichotomous items, ability estimates, item and test characteristic curves, item and test information functions (see deAyala, Koch, & Dodd, 1989; Dodd & Koch, 1985, 1986), fit statistics, standard errors that differ with ability level, and linking constants can be calculated and used for item and test evaluation.

Previous applications of the PC model to educational achievement tests and direct writing assessments suggest the model's practical viability for banking mathematics problem-

solving items (Masters, 1984); computer adaptive attitude measurement (Koch and Dodd, 1985); developing items that model vocabulary acquisition theory (Smith, 1986); equating intact forms of essay tests (Phillips, 1987); equating direct with indirect writing assessments (Phillips, Mead, & Ryan, 1983); constructing narrative writing tasks (Harris, Laan, & Mossenson, 1988); and calibrating several types of writing tasks (Pollitt & Hutchinson, 1987).

Method and Procedures

The Maryland Writing Test

The Maryland Writing Test (MWT) is administered annually to students in grades 9-12. Students must pass this test and multiple-choice tests in reading, mathematics, and citizenship in order to receive a Maryland high school diploma. The MWT is comprised of two essays written in response to a narrative and an explanatory prompt. Each prompt consists of three paragraphs which identify the elements of the writing task. Paragraph one identifies the topic, purpose, audience, and form (e.g., business letter) for the writing task; paragraph two contains planning suggestions directly relevant to the elements in paragraph one; and paragraph three restates the task in a single sentence. Table 1 contains the first paragraph from the

narrative and explanatory prompts from an annual operational edition of the MWT.

Table 1 about here

Each essay is scored independently by two trained raters who receive packets of 25 randomly sorted essays. Raters are trained in the application of Maryland's modified holistic scoring procedures which include scoring rubrics and student essays as benchmarks for each score point (see Maryland State Department of Education, 1987). Each rater assigns scores of 1, 2, 3, or 4 to each essay according to the rubric. Pairs of scores for each essay are averaged (discrepancies greater than one point are resolved by a third independent reading), and pairs of averaged essay scores are summed to produce each student's total test score. Score scales for each individual essay contain seven score points ranging from 1.0 to 4.0 in half-point units; the total test score scale contains 15 score points ranging from 1.0 to 8.0 in half-point units.

The notion of assigning partial credit to item responses that are less than fully correct or complete is directly relevant to Maryland's modified holistic scoring rubric. According to the rubric, an essay which receives a score of "2" should contain all or most of the elements and qualities of a "1" essay but only some of the elements and

qualities of a "3" essay, and so on up and down the score scale. Consistent and accurate application of the scoring rubric by raters provides the link between features of student essays and the MWT score scale. Modified holistic scores from the MWT lend themselves to calibration in the PC model, which is used in this study to calibrate individual prompts. Each prompt is accompanied by six step difficulties: one used for calculating the probability of achieving a score of 1.5 rather than 1.0; a score of 2.0 rather than 1.5; 2.5 rather than 2.0; and so forth.

Preliminary analyses indicate the feasibility of using the PC model with MWT essay data. For example, previous analyses provide evidence of unidimensionality of scores from the MWT (Baldwin, 1988). However, these analyses also indicate data-model total fit problems for some prompts, indicating the need to investigate sources of misfit and to develop guidelines for making prompt bank inclusion and exclusion decisions.

Subjects and Data Sets

Subjects are cohorts of grade 9 students who took the MWT for the first time in the years 1984-1988. These students participated in intensive writing instruction in preparation for the test and, in recent years, probably received significant writing instruction for several years prior to grade 9. Each cohort in this study comes from

randomly selected schools within a stratified sample of districts. Students in each cohort wrote a narrative and an explanatory essay for both the operational administration of the MWT and the equating administration which occurs two to three weeks prior to the operational administration.

The data for this study are averaged narrative and explanatory scores from the 1984, 1985, 1986, 1987, and 1988 annual operational administrations of the MWT and from the 1984 MWT prompts administered concurrently each year as anchor prompts in an equating study. The 1984 prompts, when administered concurrently with each annual operational administration of the MWT, are referred to here as "anchor prompts"; prompts administered as part of annual operational administrations are referred to here as "operational prompts."

Procedures

Data were analyzed using the MSTEPS program (Wright, Congdon, & Schultz, 1988) and evaluated in six steps. First, step difficulty values were estimated for the anchor prompts using data from their 1984 operational administration. This calibration establishes the anchor prompts' bank difficulty values and the bank scale. Second, each annual set of operational and anchor prompts from each of the 1985-1988 MWT administrations was calibrated separately, producing four sets of calibrations of

operational and anchor prompts. This calibration produces unadjusted step difficulty values for each annual edition of the test. At this point fit of the prompts to the PC model was evaluated. Third, linking constants were calculated for each annual edition of the test. These constants are the difference between the average anchor prompt bank difficulty value and the average anchor prompt difficulty value from the annual concurrent equating administration. Next, each edition of the MWT was linked to the bank scale using the linking constants. The result is adjusted step difficulty values for each separate prompt that are expressed on the underlying bank scale. Fifth, consistency of links across years was checked by looking for standardized step difficulty differences greater than the absolute value of 2.0 (see Masters, 1984; Masters & Evans, 1986). Last, to provide some outside comparative criterion of the results of linking prompts on a single scale, relative test difficulty as indicated by equipercentile equating procedures (see Lord & Wingersky, 1984; Phillips, 1987) currently in use with the MWT (see Ferrara, 1987) was compared to relative test difficulty as indicated by PC model step difficulties.

Results

Calibration of Essay Prompts

Table 2 contains descriptive statistics for unadjusted step difficulties and calibration standard errors for each prompt from each annual administration. (Actual values are available from the authors.)

Table 2 about here

Average step difficulties within test edition in Table 2 indicate that narrative prompts are less difficult than explanatory prompts for the students in these data sets. Differences in average step difficulties for anchor prompts range from a low of about one quarter of a logit (1985) to a high of about three quarters of a logit (1987). Differences between average step difficulties for narrative and explanatory operational prompts range from a low of about one half of a logit (1986) to a high of about one logit (1987). Contrary to this pattern, the 1985 operational narrative prompt is .128 logits more difficult than its explanatory partner. What is not evident from examining these average step difficulties is that step values are not monotonically increasing. Rather, step values follow a decrease-increase-decrease-... pattern over steps two through six for all prompts administered in all years and across almost all steps. Each decrease in difficulty

corresponds to a step from a whole number score category (i.e., 1.0, 2.0, 3.0, 4.0) to a half-point score category (1.5, 2.5, 3.5); each increase in step difficulty corresponds to a step from a half-point to a whole number score category. This pattern suggests that if an essay's quality is good enough to receive, for example, a score of three rather than two, then the essay is most likely to receive a three from both raters rather than a two from one rater and a three from the other. This pattern is consistent with the high exact agreement rates and low adjacent agreement rates achieved in all operational and equating scoring projects between 1984 and 1988, which results in low frequencies of half-point essay scores. Similarly, the probabilities of students achieving scores of 1.0, 2.0, 3.0, or 4.0 are higher than the probabilities of achieving scores of 1.5, 2.5, or 3.5. Item Characteristic Curves (ICCs) for the 1988 MWT in Figure 1 below illustrate this decrease-increase pattern of step difficulties.

Figure 1 about here

Data-model fit. Item fit statistics are indicators of how well an item fits the PC model. They provide a guide for evaluating individual items and are particularly informative in developing item banks. Fit statistics compare expected responses, given examinee ability on the underlying ability continuum that is estimated within the PC

model, to actual responses to a given item (see Masters, 1988; and Wright & Masters, 1982). It is desirable to consider excluding misfitting items from a bank because they can obscure the meaningfulness and interpretation of the latent variable represented by the bank and weaken bank coherence (Masters, 1984). However, misfitting items need not be excluded from an item bank automatically (Baghi, 1988; Masters & Evans, 1986) even though fit statistics identify items that appear not to be performing in the same way as other items. By rule of thumb, items with a fit value greater than the absolute value of 2.0 or 3.0 are considered misfitting and should be scrutinized for technical and content problems (Crocker & Algina, 1986, p. 361). Since a common cause of misfit to the PC model is calibrating small numbers of items (Wright & Masters, 1982, p. 101), we have used 3.0 as a guideline for identifying misfitting prompts. We do not ignore the potential effects of misfit on the validity of the bank scale and coherence of the bank, as will be evident later in this paper. Table 3 contains standardized weighted mean square fit values (see Wright & Masters, 1982, p. 101) for all prompts in all data sets used in this study.

Table 3 about here

Table 3 indicates that seven of the 18 calibrated prompts appear to be misfitting. Of these seven misfits,

six are either the 1984 narrative or 1984 explanatory prompt. It is noteworthy that the 1984 prompts -- the anchor prompts for this bank -- are the prompts with the most frequent data-to-model fit problems. The 1988 explanatory prompt also appears not to fit the PC model. All of the misfits are positive, suggesting that low ability examinees tended to score higher than expected on misfitting prompts and that high ability examinees tended to score lower than expected. The topic, purpose, and format of the anchor prompts are slightly different from all operational prompts and may have caused their misfit. In comparison to the other prompts, the anchor prompts impose a more structured context in which to construct written responses. This context may assist lower ability examinees in constructing essays and may restrict responses of higher ability examinees. A possible explanation for the misfit of the 1988 explanatory operational prompt is less clear cut. However, the definition in the explanatory rubric for "development of elements" was liberalized slightly just prior to the 1988 test administration, which perhaps explains its misfit.

Linking Prompts to the Prompt Bank

Linking test editions: adjusting step difficulty values. The anchor test design for this study allows linking of annual operational prompts to the bank scale: step difficulty values for each annual operational and

equating prompt pair are adjusted to the bank scale through the anchor prompts that are administered concurrently with the operational prompts. This linking contrasts with current equipercentile equating procedures used for the MWT. Under current procedures, the link between anchor prompts from a concurrent administration and bank difficulty values established from the 1984 operational administration of the anchor prompts is not provided statistically, that is, as part of the calculations of linking constants. Rather, this link is provided procedurally through prompt selection, scorer training, and scoring consistency as discussed earlier in the paper.

In this study, a linking constant is calculated for each annual test edition by subtracting from the mean bank step difficulty value, which is the average step difficulty from the 1984 operational administration of the anchor prompts, the mean step difficulty value from the concurrent administration of the anchor prompts in a given year. That is,

$$LC_k = \bar{d}_{ijA} - \bar{d}_{ijB} \quad ,$$

where k refers to a year (i.e., test edition), d-bar refers to mean step difficulties, i refers to prompts, j to steps within prompts, A to the bank step values of the anchor prompts, and B to the anchor prompt step values from an

equating administration. Means are calculated by summing step difficulty values across six steps within each prompt and across two prompts. Table 4 contains descriptive statistics for step difficulty values for all prompts that have been adjusted using linking constants. (Actual adjusted step difficulty values are available from the authors.)

Table 4 about here

Linking anchor prompts to the bank. The positive linking constants for 1985 and 1986 suggest that the anchor prompts were somewhat less difficult in those years than when they were administered as an operational test edition in 1984. Adding a constant to the step difficulty values from these tests shifts the ICCs to the right. The linking constants for 1987 and 1988 are near zero, indicating that there is little difference in the difficulty of the anchor prompts in these years from their bank values. Adjusting anchor prompts to the bank scale in this way accounts for differences in test difficulty due to ability differences in the students who participated in each annual administration and any drift in scoring standards, plus equating error.

Linking operational prompts to the bank. Adjusting step difficulty values for the operational prompts using linking constants accounts for differences in test edition

difficulty due to ability differences, scoring standard differences, and equating error, as above. It also accounts for real differences in the difficulty of the prompts in each annual edition of the test. The average adjusted step difficulty values for operational prompts in Table 4 indicate that narrative prompts selected for administration in the MWT have gotten easier since 1984 while explanatory prompts became more difficult and then easier. (Of course, examining average step difficulty values obscures changes that may have occurred at individual steps and the difficulty of reaching various score categories.) Table 5 contains differences between average operational prompt difficulty values from any two years. The differences are reported separately for narrative prompts (below the diagonal) and explanatory prompts (above the diagonal). A positive difference indicates a prompt that is more difficult on average than the prompt to which it is compared; a negative difference indicates a prompt that is easier on average. These differences are useful for discovering changes in difficulty of narrative and explanatory prompts, separate from and independent of overall test difficulty. They also could prove useful for selecting individual prompts from large calibrated banks for pairing in future operational test administrations.

Table 5 about here

Consistency of anchor prompt links. Linked or adjusted step difficulties must be consistent across years in order to avoid disturbing the coherence of a prompt bank. One indicator of consistency of links between anchor prompts across years is prompt fit. We have already commented on the misfit of the anchor prompts and the 1988 explanatory prompt. Any impact of these misfitting prompts may be detectable while examining anchor prompt link consistency. Link consistency is checked by standardizing the differences between anchor prompt step difficulties from the 1984 operational administration and each equating administration (see Masters 1984; Masters & Evans, 1986):

$$z_{ij} = \frac{d_{ijA} - d_{ijB}}{(s^2_{ijA} + s^2_{ijB})^{1/2}} ,$$

where d is a step difficulty value, i refers to prompts, j refers to steps within prompts, A to the bank step values of the anchor prompts, B to the step values of the anchor prompts from an equating administration, and the denominator is the standard error for differences between independent means. Average standardized step differences less than the absolute value of 2.0 suggest that prompts links are consistent across years. Table 6 contains information on standardized step difficulty differences for the anchor prompts from their administration as operational prompts in

1984 and from their concurrent equating administrations in 1985-1988. We report only means and standard deviations since the largest values observed are only -0.59 and 0.60.

Table 6 about here

The small differences between anchor prompt step difficulties from the 1984 administration and adjusted step values for subsequent equating administrations in Table 6 suggest that (1) the anchor prompts are operating similarly across years and cohorts, and (2) the misfit of the 1984 prompts in several administrations appears not to diminish link consistency. (However, the link consistency analysis does not shed light on the effects of misfit on relationships between anchor and operational prompts.) Link consistency appears adequate for creating a coherent bank of essay prompts.

Now that annual editions of the MWT have been linked through anchor prompts to the bank scale, we can turn our attention to the relative difficulty of annual test editions in the same data sets according to (1) equipercentile equating results, and (2) PC model adjusted step difficulties.

Test Edition Difficulty from Equipercentile and PC "Equatings"

Some outside verification could provide information useful for evaluating results from banking MWT essay prompts using the PC model. One means of verification is to compare test difficulty, as indicated by results of equipercentile equatings of total test holistic scores, to test difficulty as indicated by average adjusted step difficulty values from the PC model.

As described earlier in this paper, the equivalence of total test scores is checked each year in an equipercentile equating. Steps for ensuring this equivalence are included in procedures for developing and selecting essay prompts and for scoring essays. Equating data are collected in a sample of students in a concurrent equating administration, and scores from an annual edition of the MWT are equated to scores from essays written to the anchor prompts in this sample. These equating data sets have been used for the analysis of the PC model described in this paper. Table 7 contains raw and equated scores from equipercentile equating for three score points around the MWT cut score, and average adjusted step difficulty values for each annual edition of the MWT from 1985 to 1988.

Table 7 about here

Results from equating total holistic scores. The equipercentile equating results above produce comparisons of test edition difficulty within years. Consistency of scoring of essays written to anchor prompts allows comparisons across years. Equated scores from 1985 and 1986 corresponding to anchor test raw scores indicate that these two annual editions of the MWT were more difficult than the anchor test. This difference in test difficulty resulted in more than a half-point holistic score difference near the cut score in 1985 and less than a half-point holistic score difference in 1986. The results from 1987 and 1988 indicate that these annual editions were about equivalent in difficulty to the anchor test: the difference in holistic scores in both cases is less than a half-point.

Results from adjusting PC model step difficulties. Adjusting essay prompt step difficulties from annual test editions allows more direct comparison of test difficulty across years. Average adjusted step difficulty values in Table 7 indicate that the 1985 and 1986 editions of the MWT were more difficult than the 1984 MWT. In 1985, this difference was about one-half of a logit on average; in 1986 the difference was about three-quarters of a logit. The results indicate that the 1987 and 1988 editions of the MWT were about equivalent in difficulty, on average, with the 1984 test: the average differences in these two years are about one-tenth of a logit.

Of course, it is not possible to compare equated holistic scores directly to average adjusted step difficulty values. Equated scores correspond to score categories while step difficulties correspond to the thresholds between these score categories. However, the analysis above of total test difficulty suggests that, indeed, equipercentile equating and PC model banking procedures agree on relative test difficulty in these data sets. This result provides some tentative outside verification or assurance of the viability of the PC model for banking MWT essay prompts.

Discussion and Conclusions

The data sets used in this study were collected in five different annual cohorts of students in a context of evolving writing instruction and attitudes toward the MWT. These conditions are less than favorable for calibrating and linking a bank of essay prompts. In addition, the anchor prompts available for this study are not fully appropriate and poorly fit the PC model. Despite these problems, and because of positive results from analyses of link consistency and comparisons between equipercentile and PC "equatings," we are optimistic about the viability of the PC model for banking MWT prompts. Our optimism leads us to

speculate briefly on implications for future research, which could include designing a large-scale prompt banking project.

Implications for Future Research

Like many other large-scale essay testing programs, the Maryland Functional Testing Program would benefit from having available a large number of prompts with known levels of difficulty expressed on a single scale. However, in order to create such a bank data would have to be collected under more favorable conditions than those in this study. First, large numbers of previously administered and field tested prompts would have to be available. Second, these prompts would have to be administered to sufficient sized samples of students in large numbers of schools around the state, and within a span of several days rather than several years. Third, multiple links between prompts and a bank scale could be established through several anchor prompts. Designing prompt combinations so that individual prompts are linked to several other prompts would improve the overall coherence of the bank and help minimize linking error. Finally, misfitting and otherwise poorly performing prompts could be eliminated from the bank.

Benefits of such a prompt bank would include the ability to (1) select individual prompts to be paired for operational administration, even if those prompts had not

been calibrated as a pair; (2) eliminate annual concurrent equating data collections or at least conduct them only periodically to monitor prompt parameter stability; (3) compare directly relative difficulties of narrative and explanatory prompts during test form construction; (4) eliminate expensive annual prompt field tests, or at least, conduct them only periodically to enhance an existing prompt bank; (5) estimate prompt step difficulties using wider, more varied and representative samples of schools and students than is possible under current conditions; and (6) lessen threats to the validity of equating results caused by familiarity with anchor prompts, a current concern.

Future investigations of the viability of the PC model for establishing an essay prompt bank for the MWT and other essay testing programs should include the use of item (i.e., prompt) and test information functions (see deAyala, Dodd, & Koch, 1989; and Dodd & Koch, 1985, 1986) and investigations of differential item functioning for relevant examinee subgroups.

Limitations of This Study

We have already referred to some of the limitations of this study and the threats they pose to the validity of its results. Limitations include the misfit of the anchor prompts and the potential of this misfit for obscuring the coherence of the prompt bank; the tentativeness of the link

between operational prompts and the bank because anchor prompt data were collected during five different years; and the effects on these calibrations and links resulting from changes in writing instruction and ability over the last five years.

Importance to Other Essay Testing Programs

The results of this study add to the rapidly growing literature describing applications of the Partial Credit model to educational achievement test data. More important, this study's results should be of particular interest to other essay testing programs because (1) Usable essay prompts are difficult and expensive to generate; (2) Both pre- and post-equating essay prompts is a perennial problem which demands continued attention and attempted solutions; (3) Essay prompt equating using the PC model could prove to be feasible in situations where conventional equating is not viable, such as for typical single essay tests; and because the PC model provides a means of (4) Simultaneously calibrating essay prompts and dichotomously scored items, which facilitates test development, analysis, and equating; (5) Investigating differential item functioning (DIF; also referred to as "item bias") of essay prompts using item response theory analytic procedures similar to those used for dichotomously scored items; (6) Informed test targeting using item and test information functions; and (7) Statistically linking test editions across years.

Successful banking of essay prompts using the Partial Credit model would represent a large step toward providing satisfactory solutions to some of the current difficulties in standardized essay testing.

References

Baghi, H. (1988, March). The use of Rasch model fit statistics in selecting items for the Maryland Functional Testing Program. Unpublished manuscript.

Baldwin, J. (1988). Validating the factor structure of ratings assigned to essays: A confirmatory factor analytic approach. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt Rhinehart and Winston.

deAyala, R. J., Dodd, B. B., & Koch, W. D. (1989, March). A comparison of the graded response and partial credit models for assessing writing ability. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Dodd, B. G., & Koch, W. R. (1985). Item and scale information functions for the partial credit model. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Dodd, B. G., & Koch, W. R. (1986). Relative efficiency analyses for the partial credit model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Ferrara, S. F. (1987, April). Practical considerations in equating a direct writing assessment required for high school graduation. In A. Chafin (Chair), Using a direct writing assessment as a graduation prerequisite: Political, measurement, and instructional issues. Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC.

Goldberg, G. L., & Walker-Bartnick, L. (1988, April). Monitoring scoring standards over a rubric transition process. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Harris, J., Laan, S., & Mossenson, L. (1988). Applying partial credit analysis to the construction of narrative writing tasks. Applied Measurement in Education, 1 (4), 335-346.

Koch, W. R., & Dodd, B. G. (1985, April). Computerized adaptive attitude measurement. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." Applied Psychological Measurement, 8 (4), 453-461.

Maryland State Department of Education. (1987, June). 1987 writing supplement: Project Basic instructional guide. Baltimore, MD: Author.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G. N. (1984). Constructing an item bank using Partial Credit scoring. Journal of Educational Measurement, 21 (1), 19-32.

Masters, G. N. (1988). The analysis of partial credit scoring. Applied Measurement in Education, 1 (4), 279-297.

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. Applied Psychological Measurement, 10 (4), 355-367.

Phillips, G. W. (1987, April). Essay topic equating: Keeping the score scales constant. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Phillips, G. W., Mead, R. J., & Ryan, J. P. (1983). Use of the general polychotomous form of the Rasch model as a means of calibrating and equating indirect and direct assessments of writing achievement. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. Language Testing, 4 (1), 72-92.

Smith, R. M. (1986). Developing vocabulary items to fit a polychotomous scoring model. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Williams, P. W., & Phillips, G. W. (1984). Equating direct writing assessments. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: Mesa Press.

Wright, B. D., Congdon, R., & Schultz, M. (1988, March). MSTEPS [Computer program]. Chicago, IL: University of Chicago.

sf1401

Table 1

Illustrative Prompts from the Maryland Writing Test (First Paragraph Only)

Narrative
Suppose your teacher asks you to write about a special event you once celebrated. This event may have happened recently or long ago. Write a paragraph or more for your teacher telling about the special event you celebrated.
Explanatory
Suppose that your school newspaper is offering a prize to the student who can best explain an improvement needed in school. You know an improvement that is needed and decide to enter the contest. Write a business letter to the newspaper explaining ONE improvement you would like to make in your school.

Table 2

Descriptive Statistics for Unadjusted Step Difficulty Values for All Prompts and All Annual Test Administrations (Mean Calibration Standard Errors in Parentheses)

	Year of Administration				
	1984	1985	1986	1987	1988
Anchor prompts					
Narrative					
Mean	-.147	-.380	-.585	-.440	-.237
	(.098)	(.093)	(.132)	(.110)	(.187)
S.D.	2.257	1.618	1.339	1.934	1.762
Explanatory					
Mean	.148	-.100	-.155	.313	.338
	(.087)	(.090)	(.105)	(.083)	(.135)
S.D.	1.937	1.722	1.598	1.791	1.591
Operational prompts					
Narrative					
Mean	----	.305	.103	-.453	-.545
		(.080)	(.112)	(.108)	(.235)
S.D.	----	1.617	2.074	1.918	1.391
Explanatory					
Mean	----	.177	.633	.578	.442
		(.082)	(.090)	(.080)	(.132)
S.D.	----	1.609	1.952	1.692	1.761
Correlations between anchor and operational prompts[a]					
Narrative	--	.98	.98	.98	.95
Explanatory	--	.99	.98	.99	.95

[a]Correlations calculated separately for step difficulties from anchor and operational prompts.

Table 3

Fit Values for Anchor and Operational Prompts Administered in 1984-1988

	1984	1985	1986	1987	1988
Anchor prompts					
Narr	6.1[a]	1.4	3.3[a]	1.5	7.1[a]
Explan	4.7[a]	4.6[a]	3.6[a]	0.5	2.9
Operational prompts					
Narr	---	2.7	2.6	-0.5	2.8
Explan	---	1.5	1.7	1.2	3.3[a]

Note. Narr = narrative prompt; Explan = explanatory prompt. [a]Prompts with fit values greater than 3.0 are misfitting.

Table 4

Descriptive Statistics for Adjusted Step Difficulty Values for Annual Test Editions in 1985-1988

	Year of Administration 1985	1986	1987	1988
Operational prompts				
Narrative				
Mean	.545	.473	-.393	-.595
S.D.	1.617	2.074	1.918	1.391
Explanatory				
Mean	.417	1.003	.638	.392
S.D.	1.609	1.952	1.692	1.761
Anchor prompts				
Narrative				
Mean	-.140	-.215	-.380	-.287
S.D.	1.618	1.339	1.934	1.762
Explanatory				
Mean	.140	.215	.373	.288
S.D.	1.722	1.598	1.791	1.591
Linking constants	.241	.371	.064	-.050

Table 5

Relative Difficulties of Narrative and Explanatory Prompts Across Years

	1984[a]	1985	1986	1987	1988
1984[b]	--	-.29	-.36	-.53	-.44
1985	.69	--	.59	.22	-.03
1986	.62	-.07	--	-.37	-.61
1987	-.25	-.94	-.87	--	-.25
1988	-.45	-1.14	-1.07	-.20	--

Note. (1) Tabled values are differences between average step difficulty values; they are calculated separately for narrative and explanatory prompts. (2) Positive values indicate more difficult prompts; negative values indicate easier prompts. (3) Differences between narrative prompts appear below the diagonal; differences between explanatory prompts appear above the diagonal.
[a]The average bank step difficulty value for the explanatory anchor prompt is .148. [b]The average bank step difficulty value for the anchor narrative prompt is -.147.

Table 6

Means and Standard Deviations for Standardized Step Difficulty Differences for Annual Anchor Prompt Links

	1985	1986	1987	1988
Mean	.09	.15	.02	-.02
S.D.	.33	.33	.18	.27

Table 7

Test Edition Difficulty from PC Model and Equipercentile Equatings

	1984	1985	1986	1987	1988
Equated holistic scores corresponding to raw scores[a]					
	5.0	4.37	4.82	5.07	5.22
	5.5	4.89	5.13	5.57	5.78
	6.0	5.48	5.51	6.07	6.26
Average step difficulties from the PC model[b]					
Mean	0.00	0.48	0.74	0.13	-0.10
S.D.	2.00	1.54	1.94	1.81	1.60

Note. All calculations carried out on scores from pairs of narrative and explanatory essays that comprise total test scores.
[a]1984 scores are raw holistic scores; all others are equated using equipercentile equating. [b]Step difficulties for 1985-1988 have been adjusted to the bank scale.

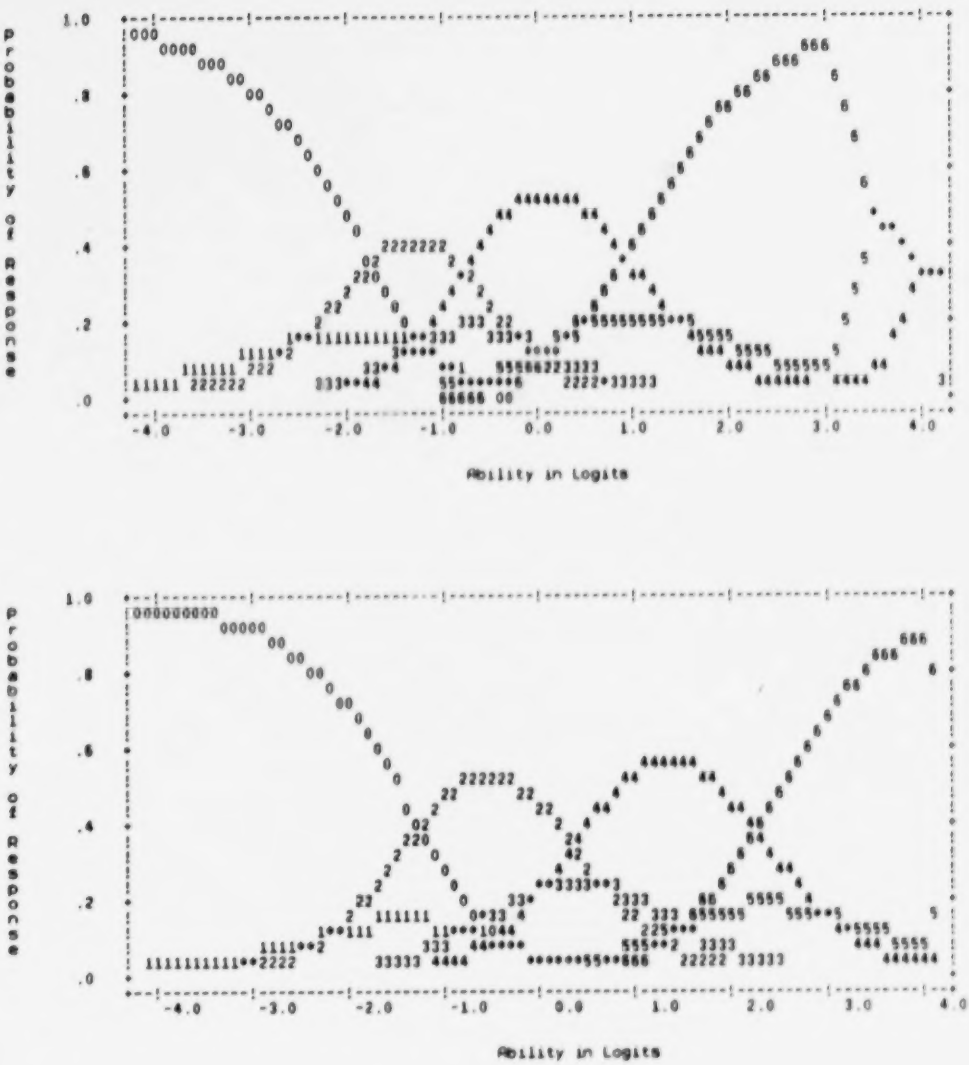

FIGURE 1
Item Characteristic Curves for 1988 Narrative Prompt (top) and Explanatory Prompt (bottom)

Note. 0=1.0; 1=1.5; 2=2.0; 3=2.5; 4=3.0; 5=3.5; 6=4.0.